# JAYPEE INSTITUTE OF INFORMATION TECHNOLOGY

## MAJOR-I
## PROJECT SYNOPSIS

## Anomaly Detection in Video Surveillance

**Submitted to -**

Mr. Amitesh

Mr. Anil Kumar Mahto

**Submitted by -**

Ayush Gupta (21103016)

Shreya Agrawal (21103028)

Kanishk Raj Mittal (21103015)

**Project Mentor:** Dr. Neetu Sardana

## 1. ABSTRACT

Anomaly detection in video surveillance is a highly developed subject that is attracting increased attention from the community because there is a great demand for intelligent systems with the capacity to automatically detect anomalous events in streaming videos. Due to this, a variety of approaches have been proposed to build an effective model that would ensure public security. There has been a variety of surveys of anomaly detection, such as of network anomaly detection, financial fraud detection, human behavioral analysis, and many more. Deep learning has been successfully applied to many aspects of computer vision. In this project, we propose to learn anomalies by exploiting both normal and abnormal videos. We aim to learn anomaly through the deep multiple instance ranking framework by leveraging weakly labeled training videos, i.e. the training labels (anomalous or normal) are at video-level instead of clip-level. In our approach, we consider normal and anomalous videos as classes and video segments as instances in multiple instance learning (MIL) and automatically learn a deep anomaly ranking model that predicts high anomaly scores for anomalous video segments.

## 2. INTRODUCTION

Video surveillance is a key application of CV used in most public and private places for observation and monitoring. Nowadays, intelligent video surveillance systems are used to detect, track, and capture a variety of realistic anomalies thereby gaining a high-level understanding of objects without human supervision. Such intelligent video surveillance systems are used in homes, offices, hospitals, malls, airports, train stations, supermarkets, schools and parking areas

Several computer vision-based studies primarily discuss aspects such as scene understanding and analysis, video analysis, anomaly/abnormality detection methods, human-object detection and tracking, activity recognition, recognition of facial expressions, urban traffic monitoring, human behavior monitoring, detection of unusual events in surveillance scenes, etc. Out of these different aspects, anomaly detection in video surveillance scenes has been discussed further in our review. Anomalies can be contextual, point, or collective. Contextual anomalies are data instances which are considered anomalous when viewed against a certain

context associated with the data instance. Point anomalies are single data instances that are different concerning others. Finally, collective anomalies are data instances that are considered anomalous when viewed with other data instances, concerning the entire dataset. Examples of an anomaly in video surveillance scenes are shown in Fig. 1; vehicles colliding, etc.
Therefore, anomaly detection can be considered as coarse level video understanding, which filters out anomalies from normal patterns. Once an anomaly is detected, it can further be categorized into one of the specific activities using classification techniques.

## 3. LITERATURE REVIEW

Anomaly detection in surveillance videos is a long-standing problem in computer vision, with a variety of approaches explored over the years. Earlier methods primarily relied on manual feature extraction and traditional machine learning models. For example, sparse-coding-based methods built dictionaries of normal behaviors and identified anomalies as events that could not be reconstructed from the dictionary. These methods showed good performance, but they often struggled in dynamic environments, where normal behaviors could vary over time, leading to high false positive rates.

Several approaches for specific anomaly detection, such as violence detection and accident detection, have also been proposed. However, these approaches are highly specialized and do not generalize well to other types of anomalies. As the complexity and variety of anomalies in real-world surveillance systems increased, more flexible models were required to handle diverse events without predefined assumptions about their nature.

More recently, deep learning techniques have been applied to the anomaly detection problem, with autoencoders being used to learn the normal behavior of scenes and detect deviations from that norm. While effective, these approaches were limited by their reliance on normal data alone, often failing to capture complex anomalous events. The use of Convolutional Neural Networks (CNNs) [4] for video action recognition also laid the foundation for modern anomaly detection systems, but these models required extensive temporal annotations, making them less practical for large-scale datasets.

The dataset introduced by Sultani et al. [1] fills a critical gap in the field by providing the largest and most realistic set of surveillance videos to date.

Compared to earlier datasets such as UCSD Ped1 and Ped2, Avenue, and UMN [5], which were small, staged, or focused on narrow types of anomalies, this new dataset offers a wide variety of real-world anomalies captured from diverse environments. This makes it a valuable resource for developing and testing more generalized and robust anomaly detection models. However, the major drawback is that the distribution of frame size is not dynamic according to the video length. So the accuracy dips to 22 and 25% for this paper. Moreover, the detection of rare events with the avoidance of false positives is a challenge.

In paper [2] behavioral attributes from video frames were extracted . Although these methods worked at the pixel level and offered better performance in complex scenes compared to trajectory-based approaches, they were limited by their inability to capture the wide range of behaviors in real-world scenarios. However, the challenge with these methods was their reliance on predefined features and their inability to adapt to previously unseen anomalies.

Autoencoders (AE), Convolutional Neural Networks (CNNs), and Long Short-Term Memory (LSTM) networks were among the first architectures applied to video anomaly detection. Autoencoders learned to reconstruct normal events with minimal errors, and any deviations from this norm were classified as anomalies. However, as Berroukham [2] et al. pointed out, not all anomalous events produced larger reconstruction errors, limiting the effectiveness of reconstruction-based methods.

Generative Adversarial Networks (GANs) have also been used for anomaly detection, leveraging their ability to generate realistic samples from normal data. GAN-based methods train on normal video frames and optical flow, allowing the system to identify anomalies by comparing actual and generated frames.

## 4. DATASET USED:

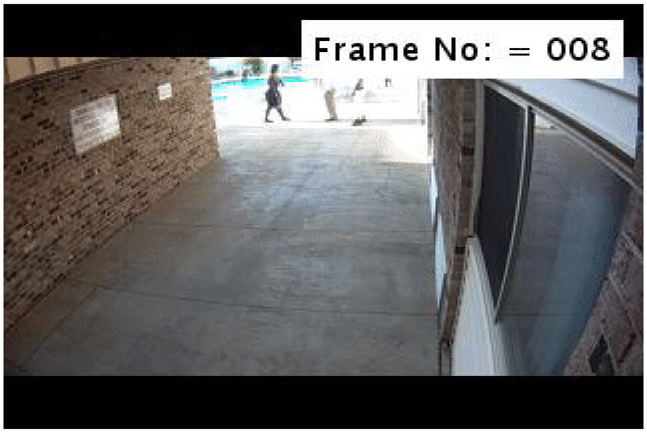
We use a dataset called UCF-Crime, to evaluate our method. It consists of long untrimmed surveillance videos that cover real-world anomalies, including:

- Assault
- Road Accident
- Explosion
- Robbery
- Shoplifting
- Normal Event

These anomalies are selected because they have a significant impact on public safety.

A short description of each anomalous event is given below

1. Assault: This event contains videos showing a sudden or violent physical attack on someone. Note that in these videos the person who is assaulted does not fight back.
2. Road Accident: This event contains videos showing traffic accidents involving vehicles, pedestrians, or cyclists.
3. Explosion: This event contains videos showing the destructive event of something blowing apart. This event does not include videos where a person intentionally sets a fire or sets off an explosion.
4. Robbery: This event contains videos showing thieves taking money unlawfully by force or threat of force. These videos do not include shootings.
5. Shoplifting: This event contains videos showing people stealing goods from a shop while posing as a shopper.
6. Normal Event: This event contains videos where no crime occurred. These videos include both indoor (such as a shopping mall) and outdoor scenes as well as day and night-time scenes.



## 5. METHODOLOGY:

One critical task in video surveillance is detecting anomalous events such as traffic accidents, crimes, or illegal activities. Generally, anomalous events rarely occur as compared to normal activities. Therefore, to alleviate the waste of labor and time, developing intelligent computer vision algorithms for automatic video anomaly detection is a pressing need. The goal of a practical anomaly detection system is to timely signal an activity that deviates from normal patterns and identify the time window of the occurring anomaly. Therefore, anomaly detection can be considered as coarse-level video understanding, which filters out anomalies from normal patterns. Once an anomaly is detected, it can further be categorized into one of the specific activities using classification techniques.

In this work, we propose an anomaly detection algorithm using weakly labeled training videos. We can easily annotate a large number of videos by only assigning video-level labels. To formulate a weakly-supervised learning approach, we resort to multiple-instance learning. Specifically, we propose to learn anomalies through a deep MIL framework by treating normal and anomalous surveillance videos as bags and short segments/clips of each video as instances in a bag. Based on training videos, we automatically learn an anomaly ranking model that predicts high anomaly scores for anomalous segments in a video. During testing, a long untrimmed video is divided into segments and fed into our deep network which assigns an anomaly score for each video segment such that an anomaly can be detected.

## 6. APPROACHES USED:

1. **VIVIT-BASE (Video Vision Transformer):** VIVIT-BASE is a transformer-based architecture designed for video processing. It leverages self-attention mechanisms to capture both spatial and temporal information from video sequences, making it effective in anomaly detection tasks by modeling long-range dependencies across frames. VIVIT-BASE excels in detecting complex patterns in videos without the need for frame-by-frame annotations, as it processes the entire video sequence holistically.

2. **I3D (Inflated 3D Convolutional Network):**
   I3D is a deep learning model that extends 2D convolutional networks into the temporal dimension using 3D convolutions. It was designed to capture spatiotemporal features in videos, making it effective for action recognition and anomaly detection. I3D processes video frames as a sequence and captures the motion between frames, identifying both the objects and the activities that define normal or abnormal behavior in surveillance footage.

3. **C3D (Convolutional 3D Network)**:
   C3D is a deep learning architecture designed to capture both spatial and temporal information in video data through 3D convolutions. Unlike traditional 2D Convolutional Neural Networks (CNNs), which process only spatial information from individual frames, C3D extends the convolution operation into the temporal domain, allowing it to learn motion and appearance features simultaneously from consecutive frames.

## 7. FUTURE SCOPE:

a) Real-Time Video Analysis with Generative AI: In future work, generative AI models can be employed for real-time video description. This involves generating textual descriptions for detected anomalies, enhancing situational awareness and enabling swift responses.

b) Future Frame Prediction: Incorporating future frame prediction techniques can improve anomaly detection systems by forecasting what should occur in the next few frames. Any deviation from predicted frames could signal an anomaly, allowing for proactive detection.

c) Real-Time Actionable Alerts (Calling Police): A more advanced feature would involve real-time alerts sent to law enforcement or security personnel during anomaly detection. Integrating video anomaly detection with IoT devices and emergency services can automate response mechanisms in high-risk scenarios.

d) Self-Supervised Learning: Future research may focus on self-supervised learning, where the model can learn representations from large amounts of unlabeled video data, further enhancing anomaly detection capabilities with reduced reliance on labeled datasets.

e) Improving Dataset Diversity: Expanding and diversifying the dataset used for training models by including new types of anomalies and covering different environments, lighting conditions, and camera angles will lead to more robust and generalized models.

## 8. REFERENCES

[1] Sultani, Waqas, Chen Chen, and Mubarak Shah. "Real-world anomaly detection in surveillance videos." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018.

[2] Berroukham, Abdelhafid, et al. "Deep learning-based methods for anomaly detection in video surveillance: a review." *Bulletin of Electrical Engineering and Informatics* 12.1 (2023): 314-327.

[3] Duong, Huu-Thanh, Viet-Tuan Le, and Vinh Truong Hoang. "Deep learning-based anomaly detection in video surveillance: A survey." *Sensors* 23.11 (2023): 5024.

[4] Nawaratne, Rashmika, et al. "Spatiotemporal anomaly detection using deep learning for real-time video surveillance." *IEEE Transactions on Industrial Informatics* 16.1 (2019): 393-402.

[5] Zhou, Joey Tianyi, et al. "Anomalynet: An anomaly detection network for video surveillance." *IEEE Transactions on Information Forensics and Security* 14.10 (2019): 2537-2550.

[6] Zhou, Joey Tianyi, et al. "Attention-driven loss for anomaly detection in video surveillance." *IEEE transactions on circuits and systems for video technology* 30.12 (2019): 4639-4647.

[7] Franklin, Ruben J., and Vidyashree Dabbagol. "Anomaly detection in videos for video surveillance applications using neural networks." *2020 Fourth International Conference on Inventive Systems and Control (ICISC)*. IEEE, 2020.

[8] Patrikar, Devashree R., and Mayur Rajaram Parate. "Anomaly detection using edge computing in video surveillance system." *International Journal of Multimedia Information Retrieval* 11.2 (2022): 85-110.